# Facilitating NASA Earth Science Data Processing Using Nebula Cloud Computing

Long Pham[1], Aijun Chen[1,2], Steven Kempler[1], Christopher Lynnes[1], Michael Theobald[3], Esfandiari Asghar[3] , Jane Campino[4], Bruce Vollmer[1]

[1]NASA Goddard Earth Sciences Data & Information Services Center (GES DISC); [2]Center for Spatial Information Science and Systems, George Mason University

[3]ADNET System Inc., [3]ADNET System Inc. and ., NASA Goddard Space Flight Center SESDA II Program Office

[4]Embry-Riddle Aeronautical University, 1413 Arkansas Road, Room 106, Andrews, MD 20762

IN41A-1398

## NASA Nebula Cloud Computing Platform


Cloud Computing has been implemented in several commercial arenas. The NASA Nebula Cloud Computing platform is an Infrastructure as a Service (**IaaS**) built in 2008 at NASA Ames Research Center and 2010 at GSFC. Nebula is an open source Cloud platform intended to:

a) Make NASA realize significant cost savings through efficient resource utilization, reduced energy consumption, and reduced labor costs.

b) Provide an easier way for NASA scientists and researchers to efficiently explore and share large and complex data sets.

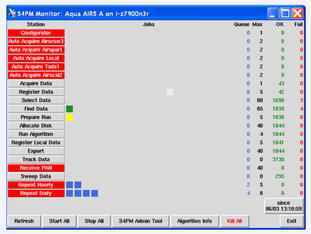
c) Allow customers to provision, manage, and decommission computing capabilities on an as-needed bases. NASA Nebula: http://nebula.nasa.gov/

## Cloud Computing Projects at NASA GES DISC

NASA GES DISC has been evaluating the feasibility and suitability of migrating GES DISC's applications to Nebula platform by porting the following projects:

**a) Using Nebula Cloud to run scientific data processing infrastructure**

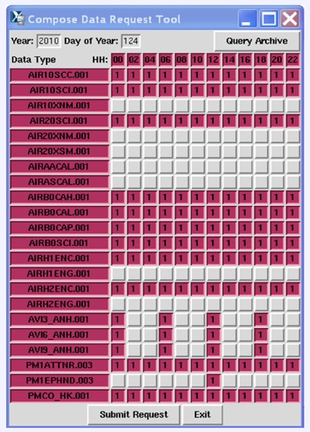
**S4PM** is an open source data processing infrastructure. Based on S4PM, scientific data processing algorithms can be run to efficiently process large volumes of satellite data.




S4PM Data Processing Monitor GUI

**b) Using Nebula Cloud to run scientific data processing workflow**

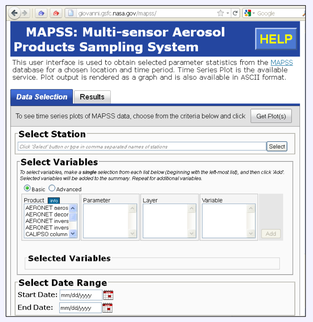
The Atmospheric Infrared Sounder (AIRS) focuses on supporting climate research and improving weather forecasting. Based on S4PM, the **AIRS Level 1 & Level 2 algorithms workflow,** consisting of many of sub-algorithms (executables), processes large volumes of AIRS Level 0 data to produce Level 1 data as intermediate results, and finally outputs Level 2 data products.




AIRS L1/2 algorithm workflow Data Request Composing Interface

**c) Porting a Web-based scientific data processing application to Nebula Cloud**

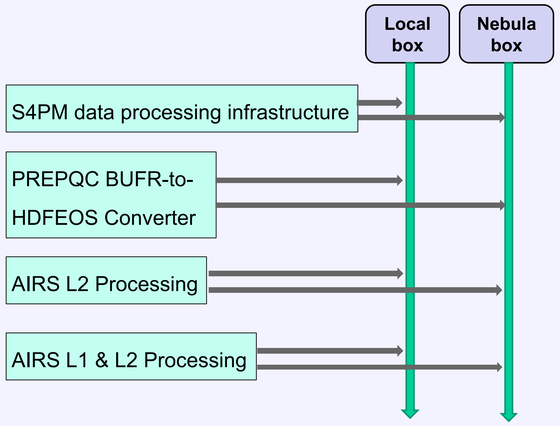
**Giovanni** is a Web-based application which offer online visualization and analysis of vast amounts of Earth science data. The **Giovanni MAPSS** (Multi-sensor Aerosol Products Sampling System) portal focuses on visualizing aerosol relationships among ground-based data and satellite data.




Giovanni MAPPS instance running in Nebula

## Migrating AIRS L1/L2 Algorithms Workflow

Running S4PM requires installation of auxiliary packages. The AIRS L1/L2 algorithm workflow runs based on S4PM infrastructure and involves quite a few libraries, e.g. HDF, *sdptk*, and basic data, e.g. DEM, MODIS, AVHRR. Migrating it can be time-consuming. The diagram at right shows the procedures for pre-installation and testing of S4PM and AIRS algorithms first on the local box, then the Nebula box.



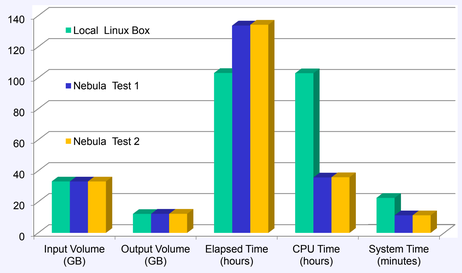

## Performance Comparison between Nebula & Local

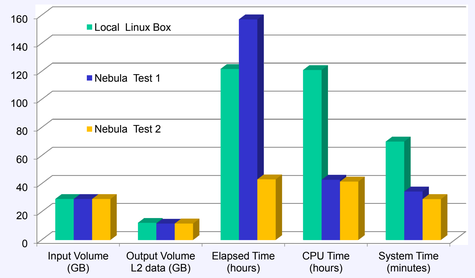
| Two days (2010.123-124) | Local Linux Box | Nebula Test 1 | Nebula Test 2 |
|---|---|---|---|
| Input Volume (GB) | 33.1 | 33.1 | 33.1 |
| Output Volume (GB) | 12.16 | 12.2 | 12.2 |
| Elapsed Time (h) | 103.05 | 133.60 | 134.13 |
| **CPU Time (h)** | **102.90** | **35.67** | **35.80** |
| System Time (m) | 22.47 | 11.27 | 11.27 |

**Input Data (L1B):** Calibrated and geolocated radiance in physical units, e.g. brightness temperature in Kelvin (K).

| Two days (2010.123-124) | Local Linux Box | Nebula Test 1 | Nebula Test 2 |
|---|---|---|---|
| Input Volume (GB) | 29.11 | 29.11 | 29.11 |
| Output Volume L2 data (GB) | **12.14** | **11.61** | **11.64** |
| Elapsed Time (h) | 121.70 | 157.00 | 43.11 |
| CPU Time (h) | **120.98** | **42.80** | **41.52** |
| System Time (m) | 70.02 | 34.43 | 29.04 |

**Input Data (L0):** Raw data from AIRS, AMSU-A1, AMSU-A2 instruments, and data about the spacecraft.




**Output Data (L2):** Retrieved physical variables, e.g. temperature, humidity and ozone profiles, total precipitable water, cloud top height.

## Cost Comparison between Nebula & Local

| AIRS L1/L2 Processing | Total cost for 3 – 5 years | Average cost | Notes |
|---|---|---|---|
| At Nebula box | $8,159.40 - $13,599.00 | $0.47/GB | Not include instance setting up, test, idle. |
| At Local box | $10,012.29 - $16,687.14 | $0.57/GB | Update, maintenance, administration. |
| Actual AIRS Processing | $80566.28–$134,277.14 | $4.60/GB | Update, maintenance, administration. |

Using Nebula Cloud to run scientific data processing is **faster**, with **much lower cost**, compared with the current data processing system.

## Hardware Performance Analysis

| | Local Linux box | Nebula virtual Linux Box |
|---|---|---|
| Hardware | DELL PowerEdge 6800 with Dual-Core Xeon Processor 7100 series / 4 CPU | DELL PowerEdge c2100 with Quad-Core Xeon Processor 5500 series / 2 CPU (2 c2100 offers 4 CPU virtually) |
| CPU (GHz) | 4 * 3.16 (8 cores available) | 4 * 2.8 (16 cores available) |
| RAM (GB) | 16 | 8 |
| Storage | 11TB | 300GB (200GB in default) |
| CPU Microarchitecture | 65nm NetBurst | 45nm Nehalem/32nm Westmere |

| | NetBurst Microarchitecture (Local Box) | Nehalem/Westmere Microarchitecture (Nebula) |
|---|---|---|
| Cache L3 | N/A | 2 MB/core |
| FSB | Dual Independent 800MHz | QPI=6.4GT/s (QuickPath Interconnects) |
| Memory | DDR-2 400 ECC SDRAM (double channel) | DDR-3 (triple channel) |

**Netburst** (65nm) --> **Core** (65nm) /**Penryn** (45nm) --> **Nehalem** (45nm)/**Westmere** (32nm)

Core = *2.5 x* NetBurst
Penryn = 1.8 x Core
Nehalem/Westmere=*(1.2-2.0) x* Penryn

⇒ Nehalem/Westmere=*(5.4-9.0) x* NetBurst

## Advantages, Lessons, and Challenges

### Advantages of NASA Nebula Cloud Platform:

- ❖User friendly interface, access to and management of Nebula resources; dashboard & Euca2ools.
- ❖Better performance compared to local box
- ❖Lower cost (only pay for used time and resources)
- ❖Scalability, on-demand provisioning of resources in near real-time ,and no user involvement for peak loads
- ❖Cloning, simple bundling process to save a modified/ improved image.
- ❖An excellent feature to maintain, back up, and mirror the systems; hence, increased reliability.
- ❖Knowledge base, including detailed instructions, tutorial, and FAQ.

### Lessons Learned:

- ❖ Bundle early, bundle and backup often !
- ❖ Take detailed notes:
  - Record each step taken to launch and install missing and required software packages.
- ❖ Acquire SA assistance
- ❖ Use same directory structure
- ❖ Use Euca2ools
- ❖ Expect the process to be time-consuming

### Challenges Faced:

- ❖ Stability – e.g. portals are not stable, network (FTP/wget) is slow and not stable.
- ❖ Underdeveloped (e.g. Object Store) managing and monitoring tools.
- ❖ Bare-bones images, wrong location of attached volumes, some defects in the bundled images.
- ❖ Gaps in Knowledge Base.
- ❖ Size Limitation, e.g. limited size of volume, at most 16 cores
- ❖ Commercial Software installation and licenses.

## Future: Making operational system at Nebula

a) Migrate more of GES DISC's applications/portals, e.g. Giovanni portals, to the Nebula Cloud platform;

b) Making mature migrated applications operational on the Nebula Cloud platform.

c) Testing some commercial Cloud applications designed for government, e.g. Amazon GovCloud.

**Acknowledgements:** Authors affiliated with **Center for Spatial Information Science and Systems (CSISS), George Mason University** have a cooperative agreement with GES DISC (Agreement No.: NNX06AD35A, Center Director: **Dr. Liping Di**).